A1  
 Perfect. And then I would first ask you to briefly introduce yourself again.

I16  
 Yes, hello, I am XXXXXXX. I am also a computer scientist, I actually studied at XX XXXXXXXX. And since 2018 I have had the company XXXXXXXXXX XXXXXXX XXXXXXXXXXXXX. We deal with digitalization for real people and of course in the last two years we have also increasingly focused on AI. Throughout my career I have had many touchpoints with it, since I spent a long time in mechanical engineering where this was a big topic. The whole field of machine learning. Before that I was in medicine, where it was also a huge topic, especially in image processing. And for me it is now also a very big topic, because we say we bring together very different aspects of digitalization. How can we design this in a way that is humane and acceptable for people? What is technologically possible for that and also what areas of application and what technologies exist at all? And for us the topic of AI in small and medium-sized enterprises is, for example, super exciting. And of course also the question of what kind of security concerns it brings with it, which exist for good reason. And therefore also the topic of trustworthiness and reliability in AI systems. That is something I have also been working on very intensively in recent months.

A1  
 Okay. [Bad connection] not so much working with AI yourself, I mean in a value-neutral sense, but rather exchanging ideas with people about AI and advising and supporting companies a bit. Did I get that right somehow?

I16  
 Yes, we also do very clear model projects where we say, okay, we train neural networks ourselves for specific fields of application. But very often for us it is really consulting or even first raising awareness. So for over a year I have been regularly traveling through various business networks, giving talks on the topic of how AI actually works, how deep learning works, how transformer architectures work and why I as a managing director need to know that if I want to apply it meaningfully, even if I am not interested in all the technical details. And I find that very enlightening, on the one hand. So people like it very much to see, ah okay, this is not rocket science, someone can explain this to me at “Sendung mit der Maus” level. But it is also important that people realize this is not automatic. It is not that it just happens without work and you snap your fingers and it runs. And I think that is also a central realization that people take home and usually the right projects then grow out of it, where you ask what can we actually do with reasonable effort, where does it really get us further. And can we trust the whole thing. That is also a very, very big topic.

A1  
 Yes, definitely. But cool. So also really in depth sometimes, prototypically developing solutions, training a few models, working something out. Whereas the operationalization is probably more with the companies again. So you basically just make the first move. But it is definitely cool that you are both in depth and also a bit high-level. Because yes, I know that from my work as well. I am also involved in an industrial project and the way people approach AI and what they think of it is sometimes very different. And what would you say, when you think about your everyday processes or also the processes you accompany, which are then also your everyday processes, are the biggest challenges in the context of AI?

I16  
 So what I actually stumble across most often is that people try to take the second or third step before the first. I think if you want to deal with AI, even as a non-technical person, then there is a certain basic knowledge that you should have. About digitalization, because there needs to be a [Bad connection] otherwise there is no AI competence, and many people try to skip that now and say I do not have to learn all that anymore, AI does it. And that is just not how it works. For example, I do a lot of consulting, especially in HR, where the topic is actually very well received. Things like evaluating applications, pre-sorting applicants, optimally checking where people can be further developed, which trainings they should do. And for example the topic of data protection and privacy is very often ignored there. And also bias effects, which unfortunately inevitably occur in AI, are at first simply ignored because people do not have that on their radar at all. Because the statistical basic understanding is missing, because the technical basic understanding is missing. And that is something where I notice that this always has to be built up first before it can move forward.

A1  
 So fairness is also of course a big topic in AI research, it always somehow comes back to our data that are collected statistically, and those are full of biases because we as a society and as individuals have and reproduce biases. And in the end you always need people who even recognize these biases in the first place, and then you can try to correct them, also in statistical datasets and models. But yes, I can immediately believe that people who do not have this technical knowledge are not even aware of it and just think, well, a preselection is great, it saves us work. But it potentially introduces strong bias. Can you think of something where you would say, that would simplify the work? More generally, that would, well, we are talking a lot about knowledge transfer in this context. Where do you feel we as ML researchers or AI researchers could get better to simplify such processes?

I16  
 I think the topic has to move much more into the popular science area. There are already great initiatives, for example I am thinking of an AI Campus that is really fantastic and does a lot of educational work. But I think it also has to happen in other ways. Especially now, where we have progress timelines in technology that are absurdly fast. In that situation, it is not very helpful to just communicate absolute basic knowledge if the practical systems are already three generations ahead. I see a lot of potential there, that academia could also support by preparing things in a way that allows people from outside the field to even understand what is happening, how it can be assessed, and how it can be assessed in a meaningful way. Because on the one hand there are people panicking and already seeing the Terminators standing in front of their own doors. And on the other hand there are people saying we just have to move forward with progress and ignore all risks, and I think both positions are completely wrong. And this does not really get through to society. I also work a lot in schools. I have been XXXXXXXXXXXXXXXX for many, many years. I am on the XXXX XXX XXX XXXXXXXXX XXXXXXX here in Germany for the national competition. And you can see it there too. AI is a topic that schools or even universities can no longer ignore. And when I see, for example, the case some months ago at LMU with the student who wanted to enroll in the AI master program and was rejected because he had written his application with ChatGPT, that shows me that there are still quite a few discrepancies, even within academia, where administration is maybe years behind technological progress. So I think there is a lot that can be done through education and by doing this in a language and format that can actually reach a broad audience.

A1  
 Yes, I completely agree, I think all of us in the team agree. That is also part of our motivation and why it has to happen in an interdisciplinary way. Can you make that a bit more concrete? Because educational work in simple language is still a relatively general idea. Do you have specific ideas or measures in mind?

I16  
 Well, basically it starts right at the bottom when you talk about probability-theoretical terms. People are just really terrible intuitive statisticians and estimate probabilities completely wrongly. That means you have to approach things carefully and observe them. And you also have to understand that AI systems are stochastic systems and therefore give stochastic answers, and that makes some things possible. But not everything, and especially not everything with absolute certainty. And that is a basic understanding that has to be there first, so that people can understand that there are also limits. For example, the law of large numbers. That has to be fulfilled. It is no use if you as an HR department say, I will just throw in 30 applications and now the system is trained. That does not work. And those are the kinds of things I would say people have to be aware of. And then also be aware of what you concretely need to know given what you are constantly bombarded with nowadays. Whether as parents interested in their children’s schooling, as employees in a company, or as business owners who have to deal with the topic. They all get hit with marketing material from every side. And it is always bigger, faster, further, something new every week. Most of the changes are only in the details, and most of the time they are just benchmarks that have been nicely polished. And the actual state, what can we do with AI and what can we not do, what will be possible in the foreseeable future and what will not, what have we understood and what have we not yet understood as researchers in the field. I think these are things that need to be explained, to show people that this is not far away from their own thinking, they were just missing essential information. And that what researchers are doing is not science fiction at all, at most what the marketing departments of AI companies are telling us is science fiction. But I can maybe distinguish that myself, because I learned that it works on a statistical basis, that it has certain limitations. I learned roughly how neural networks function. I have an idea of how that works. I have an idea of what can be done with it and what cannot. What data are collected and what are not. And I think there is still a lot of room for improvement there.

A1  
 Yes, I think so too. It really is basic knowledge in the end that is missing in many places. There is, I am probably not telling you anything new here, the gambler’s fallacy, that people cannot realistically assess random experiments like a coin toss and then think otherwise. The whole concept behind gambling really, why people go to casinos, because they think they can actually win there. And of course they can, but with a very, very, very low statistical chance. And most people lose money. And that shows that we as a population are not really trained in this yet. Yes, totally true. We have also come up with something that could perhaps make knowledge transfer a bit easier and better. For that I would quickly share my screen. I just have to quickly start the PowerPoint.

A4  
 Before we move on to the next step, could I maybe ask again what exactly your role is? I am not sure if I have not understood it yet or if we have not really talked about it in detail. What exactly do you do, what are your connections to AI?

I16  
 So, I generally try first of all to build a bridge between information technology and normal people in companies, because I speak both languages and can act as a translator, and I can also take away people’s fear of even dealing with the topic of technology. AI is really something that has just been added on top in recent years. It had already been present much, much longer before, because as I said, in medicine, in mechanical engineering we have been using it for many, many years, but back then it was more of a playground for the engineers and the rest did not have to deal with it. Suddenly everybody has to deal with it. And suddenly everyone is also being told that if you do not deal with it, you will lose your job in a few years. And that of course fuels completely new fears, and that is where I also come in as a consultant. As a coach for teams, for individuals who say they do not want to stand completely blind in front of the topic. So a lot of what I do there is consulting work. And later on, in fact, also on the technical level as a software architect, supporting technical teams that then have to implement these things concretely, creating prototypes with which you can test whether things even work the way you imagine them to. Those are the kinds of topics I deal with.

A1  
 So originally it was actually about digitality in general, and now in the last two years, when AI has really become market-ready, like you said, regression is something we have already been doing for 20 years in industry, but now that you can just buy chatbots and so on or that there are completely new possibilities for solutions on the market, the AI topic is now coming on top as well?

I16  
 Correct.

A1  
 Good. Yes, exactly. As I said, we are also dealing with AI and in particular with knowledge transfer and we have also come up with something for that. You know this because you signed up for the study. For this I will now really share my screen, as I just announced. Can you just give us an immediate reaction to what you see?

I16  
 This is a label with quite a lot of information. To me it looks a bit like I can judge quite well how energy-intensive the whole thing is, how reliable the whole thing is, how fast the whole thing is. And of course I see that it is A, green. That means it is the best I can get. That is always nice, of course.

A1  
 Exactly. So we have different metrics at the bottom, and at the top some kind of overall rating on a scale. Are the terms familiar to you so far or are there things you cannot place at all?

I16  
 Well, with the corrupted robustness I cannot do that much in this form.

A1  
 Very good. You are not the first one. Exactly. Let me maybe explain that briefly. So you probably know accuracy. That is the thing you always want to maximize in classification. With corrupted robustness the idea is that the test data that was used for the accuracy calculation gets practically modified or post-processed. Not in a targeted way, but rather by applying different filters that change it in some way, in order to see how robust the prediction quality is when you, for example, just add some normal Gaussian noise. But also maybe more specific things, like some perspective distortions or a little bit of fog, or something that could realistically occur in a camera image, in order to estimate how much this robustness is maybe overfitted to this dataset and breaks down the moment you throw in completely different or even just slightly different data. Exactly. Then there is also... to digress a little, there is also a direction in machine learning that deals with adversarial training, where people really say they want to deliberately use that already during training to make models more robust. And a further step is when you even carry out targeted attacks and, for example, use the gradients from a network to trick the model with as little change as possible, which of course is also a very relevant research field in terms of security. So let us assume you do not have full control over the data, but there is someone who could, for example, stick something onto road signs. Sure, if you black out the whole sign, then you still recognize it as a stop sign, because at least it has the shape. But the question is, what is it that the model can withstand? Exactly. The terms above, MobileNet, infer, ImageNet. Are they familiar?

I16  
 I would guess MobileNet V3 Small. That is the model we are specifically talking about here. ImageNet is probably the model family, but I am not entirely sure.

A1  
 ImageNet is an image dataset for image classification. It is one of the largest datasets for image classification and it has been heavily mined. There are lots of models that have been trained on or for ImageNet. Exactly. And by inference we simply mean that the model is just being evaluated. So it is about the practical downstream use of this model. It is not about retraining it. That is also why there are no metrics shown below that relate to training. Exactly.

I16  
 And what I see there in between, A100x8, that is simply hardware and ultimately software that the model is running on, right?

A1  
 Exactly right. I always call that the environment in which the model is being operated. Because in fact all the metrics displayed below absolutely depend on the scenario in which the model is being used. In particular, of course, runtime and energy consumption. But it is also the case that if you install a different version of TensorFlow, the quality can fluctuate slightly. Because some algorithms are implemented differently. Exactly. Do you see any connection to your own work? Do you have the feeling that this is something that could be useful in your professional practice in some way?

I16  
 Yes and no. I think it is definitely useful once you get to the stage of the selection process. So, what exactly do we want to use as a model? What is relevant for us? I think with such a label you have very good opportunities to say, what is important to us? How do we see that reflected here? And then we can make a good choice. And the problem that, in my opinion, strongly prevails in practice at the moment is first of all to even find the right use case. And I cannot just pick a model if I do not know my use case at all. And this does not really help me at that stage. But I think once you are beyond that, then this could be very good, if you say, I have such a catalog to follow. Where I can maybe also search for certain criteria. What is important to me? What minimum requirements do I have in certain places? Then something like this is very practical.

A1  
 Yes, that is exactly what it is meant for. That is also why we wrote in the interview study that we want to sit down especially with developers for these conversations. But it is of course also super helpful to get feedback from other people. The design is a bit inspired by familiar labeling formats, for example the energy labels from the EU. This scale looks a bit like the Nutri-Score. So these are things that are somewhat familiar to everyone. At the same time, we are not in a situation where everyone can just go to the supermarket around the corner and buy an AI. Everyone may buy a washing machine somewhere, but nobody just buys AI like that. Instead, these are decision processes that primarily lie with companies. And whether a product has AI in it or not, that only matters to the end user to a limited extent. They can only influence that to a limited extent as well. Exactly, but our idea is that such a label could maybe simplify communication in the development and provision of AI products. Exactly. Another advantage is that you can also create a certain form of comparability. That is why here is a second label as well. Could you maybe briefly summarize what you see?

I16  
 Well, if I look at it now like this, I would say I naturally first make a comparison of what performs better overall, because I can immediately estimate that well with the scale on top. And if I am then interested in what matters to me, then of course I look down and see, okay, if I really want to have good accuracy, then I would naturally choose the one on the left. And then I might have to live with the fact that the power consumption is higher. And that is then a concrete decision I have to make. But that is actually already something that would really make it possible to compare things well.

A1  
 Also somehow a bit difficult at the moment, because you then somehow have to read very, very long papers or so, in which the models are discussed in depth. And in particular, there is actually no best model. That is a bit counterintuitive. On the upper scale it looks like there is a best model. But that is always only given your own priorities. So if you say, I want the fastest model or the most accurate model or the most robust model, then the ranking changes completely. And that is also something that many are actually not aware of. Companies of course always want the best. And the best simply means for them the one with which we earn the most money. But that then has to be translated into technical metrics. And that is not trivial. And that means you have to weigh the advantages and disadvantages better somehow. That is also a bit of the motivation behind my research field. Resource-aware machine learning. We have now done ten years of research and always presented the new best model, the new best method. But actually that only means we have used twice as much electricity and now have better results. Then you can also critically question whether it is really better or just different. Can you think of anything where you would now say, from your work, from your expertise, that this should be solved differently, also on the presentation level?

I16  
 Yes, so what does not really become clear to me is how the value goes into the overall scoring. Yes, well. Personally, I would also say, I deal with many companies that think much more sustainably, that a high rating of energy consumption is definitely important. Runtime for example is something where I would say that is a nice metric, but it depends very, very heavily on the purpose of use how relevant it is. And whether it really has to weigh heavily in the rating or not, I do not know. And with accuracy it is exactly the same. It can be that very low accuracy values are good enough for me. But it can also be that I have requirements that put immense importance on accuracy really being given. And with robustness it is actually the same. Where I would say, a fog like that does not really help me. I even think it could quickly lead you down the wrong path at first. Because it does not really correlate with the purpose of use.

A1  
 Yes, so that definitely has to be decided very individually. Here, just to explain it once, it is like this: we actually even have a few more metrics than the four we show below. For example, we also looked at Top-5-Accuracy. Of course that is strongly correlated with Top-1-Accuracy. The question there is whether the correct class is among the five most likely predicted classes. That naturally gives you another target metric, so to speak. But also one that is relevant for many companies, especially when you think of recommendation systems. They usually do not want just one recommendation, but the Top-5 recommendations. That means depending on the use case, the metrics below also differ, and also their relevance. And in our case we just said, we have metrics that somehow relate to resource consumption, for example time as a resource or energy. And we have metrics that relate to quality. Within these groups we weight everything equally and we also weight the two groups equally. That is the very simple ranking we are doing here. But in fact, also in the papers I have written, the importance of interactivity in such labeling procedures or a framework that would enable that is relatively central and important. And maybe you are in a runtime-constrained setting, where you say, we want to deliver solutions or answers every x seconds. Then runtime suddenly becomes very important, or at least that it stays under that threshold. Exactly.

I16  
 Yes, those are exactly the cases where I think, if a bank is now running a real-time analysis, a fraud analysis, then I need very, very high runtimes. And then I might accept that this consumes a lot of energy and that I get many false positive results, while others say, I need absolute accuracy now. And for that it can be slow, because I let it run on our data lake at night and it pulls out 10,000 datasets. And if it takes all night, I do not care. So in the end it really depends strongly on the use case that comes into play there.

A1  
 Yes, definitely. Exactly. In your work, when you talk about AI and machine learning or also support companies in that area, what form of information or communication do you use there?

I16  
 When I support them?

A1  
 Exactly, where do you get your information on the topic from, when you educate yourself or also when you want to inform others? Like, this is something you can read up on again.

I16  
 So I have several agendas and newsletters, where I basically get sent a few interesting current papers several times a week. And that is an important source of information for me. There are several people who also deal with this topic, from whom I also subscribed to newsletters, subscribed to blogs, and get informed through that. About what is going on and what they are working on. Of course, also company websites, especially in the LLM field, really checking regularly what OpenAI says, what Anthropic says about the current things they are working on. I think that is also an important topic. And then also in other areas, in cross-sectional areas, to look at what the industry associations are saying at the moment. When it comes to security, what does the BSI say about it? These are also things I check regularly to see what is happening in that field at all.

A1  
 Mhm. What if you now have something very concrete, so a paper is already pretty concrete. The other things are of course also completely valid. But that is more about keeping an overview. When it comes to concrete models or very specific solutions, is there anything else besides papers where you inform yourself?

I16  
 Yes, so those are the kinds of things where I think, okay, has someone in a certain [bad connection] curated lists, model lists, where you do not have to click through half of HuggingFace to get to know the common models in one area, but someone else has already done it. And that is really nice in this field nowadays, that almost everything has already been done by someone else and with that you can actually get pretty far. Otherwise, of course, keyword search is also really an important topic when you know very specifically what you are doing. For example, last year I had a project that was quite focused on text-to-speech and speech synthesis, which has become an incredibly broad field by now. That was relatively problematic, finding things there, because what you find through search engines often depends a lot on what is currently trending, and that clutters up your search results when you want to look something up. What really interested me there were open-source models that are freely usable, that also work with relatively few resources. That was an analysis that was relatively difficult, I must honestly say.

A1  
 In the end, did you find something that somehow served your purpose?

I16  
 No. Well, I found things that were good enough for a prototypical use at first. But the big challenge was that it was about German as a language in text-to-speech, and that makes it especially difficult, because you cannot just say, well, then I will take something like Tacotron or anything derived from it and it will work. I think in the end we went with the Solero models because they still delivered relatively good results, but if I were to look again now, I would probably already find completely different things.

A1  
 Yes, definitely. In the context of the study we also looked at whether we could at least partly automate transcription and we are now using a branch of Whisper. Whisper is an LLM for that. A branch of it, developed by RWTH Aachen, that can also run locally somehow. That was of course a criterion for us for the evaluation. We are actually quite satisfied with it. Yes, XXXXX can say more about that. She goes through all the transcripts again and corrects everything the AI did not get right. But overall it already works, right?

A4  
 Yes, sometimes a few paragraphs are missing or so, but apart from that it actually works really well. I am impressed, because some words are recognized where I think I would never have understood that, and then other words that seem obvious are just wrong. But apart from that it makes the work 100% easier.

A1  
 Yes, that is again the topic of robustness at that point. It is just not completely deterministic which pronunciation works and which does not. Well, in the framework of our study we also asked ourselves a bit what alternatives to labeling are currently on the market, what is being used, what is the current workflow for people. Those are the things we found. So, on the left for example, quite classic, there is also a paper on MobileNet, that is about 40 pages you can look at if you really want to understand in depth how this model works. On the right is the corresponding model card as listed on Hugging Face. There is this model card format, which originally came from the Google paper, and by now it has also arrived there, where you just summarize the important KPIs and information. Here measured on a Galaxy from Samsung. That means MobileNet is of course also designed as a model especially for use on mobile devices, because it needs relatively few parameters or relatively little memory and is relatively fast. Exactly, and fitting to that we also have the overview from Papers with Code. That is a site that is very practical for everyone who really wants to get to the code level, because in the end anyone can voluntarily link their paper with their code there and then you get for all listed models for example also a list of all papers in which this model is used and so on. It is kind of an open database on that topic. Similarly open are blog posts. By now there are also corresponding blog posts on Medium or Towards Data Science for all major technologies. Of course, with such things there is always the risk that anyone can report anything and you have basically no trustworthiness behind it. On the left we also have the current Keras documentation, that is how the model is available and can be used in TensorFlow Keras. Again, especially relevant for developers because you also have the software interface directly there. Some KPIs are usually reported there too. And on the right a factsheet, not for MobileNet actually, because this factsheet format from IBM is really only used for IBM products and IBM models. It is also patented and so on, but it is actually very similar to model cards. If you now look at this comparison, maybe you can go into more detail about where you see advantages and disadvantages and how this comparison of these different things looks to you.

I16  
 Yeah, well. So with papers it is of course the case that they are especially relevant when it goes into depth, although I would also say that quickly putting a paper into an LLM and generating a summary is also part of my standard workflow, so that I can first see does it make sense to deal with it and read it or can I just ignore it right away? That way you can also save time. But still, usually this is something that is relatively complex. Model cards are something I also use regularly. The nice thing there is that they are classified, which means if I have a specific problem and I know what classifications I need to look for, then they are quite good for identifying suitable solutions. With the other things, well the IBM thing I don’t know, I can’t really judge how valuable that is. For the developer documentation I would say that is more for practical application, how it is used in a larger context, I see that as important in a later process step (?). Blog posts I find very important, but for me they are more about general information. I would not search for specific models through blog posts. I have, for example, subscribed to something like *Towards Data Science* and I do read things there, but I usually skip those articles that are very specific, since I often have the impression they are quite biased. In that case I would rather look directly at the related papers to get an overview. The question is always the question I actually have. Do I need a solution? Then I would say at that point going through a model card is the first step that makes sense, if you have some kind of database like what we see up there in the top right. That could of course also be very helpful, but with Models with Code it obviously very much depends on whether the researchers or developers actually provide their material there, so those are also pieces of information that can be relatively incomplete. Once I roughly know what could be relevant, then I would say things like papers or the additional documentation, which might also be linked further down in the model card, come into play to determine whether something is suitable or not. And that would also be exactly the point where I would say an AI label would make sense, because that could help make a good preselection. If you now have five different things that could be suitable, and I would say five is a very low number, because the number of models available for specific purposes keeps growing, then something like that can help to reduce the number of potential models relatively quickly. In my opinion that would be a good point of use, and only afterwards, once that has happened, would things like the developer documentation actually play a role.

A1  
 Yeah, as you just said yourself, decision processes, there are more and more models that could be suitable for that. That is of course in principle correct, but at the moment it is also partly the case, as we noticed in the interviews, that the understanding of what actually is an AI model is very different. On the one hand there is this whole area of large image models or even larger language models, and of course that is where most of the attention is focused. Very often when working with companies, and you already mentioned this, they often have much smaller datasets where even a neural network is usually completely overpowered, because you end up with more parameters than training points. Sure, it can learn everything, but it no longer generalizes at all because it can just memorize everything. How often do you actually experience such… Do you often encounter such decision processes?

I16 00:04:04  
 Yes, quite often, because here again this general lack of knowledge plays a role. People then say, hey, I saw this really great white paper or I heard some signal from OpenAI and they say you can do this and that and everything with it, and you just need to call the API once and all problems are solved. And then when you look at the actual problem you realize, well, that is a very specific case. [Bad connection]… It would be better to apply a simple model. I have often seen people trying to solve something with neural networks where K-means clustering would have been completely sufficient and the more reasonable choice, because, as you just said, when the dataset is so small you immediately get overfitting and in the end you recognize nothing except the cases you have hand-trained, and that is of course bad. It is mainly about educating people and telling them, do not shoot sparrows with cannons, look at what you are actually doing. We can skip all that research work of which models are really good if your use case is so narrow that we can probably get to the goal much faster with a self-trained model or with a classic ML algorithm that is much simpler.

A1  
 Yes, exactly, I see it the same way. I just wanted to follow up on that, because I think this… There are always so many different options. You can do almost everything with Random Forests or with K-means or with neural networks, and then neural networks are not all the same either, since there are many different layer types. You can also construct very simple, interpretable neural networks with just a few hidden neurons and that is it. Then you still have a manageable number of parameters and every decision can be traced back to some extent. But what deep learning is known for is simply huge datasets and incredibly complex tasks that are then solved.

I16  
 Yes, but that is really part of the problem nowadays, that people say, hey, with AI I can solve everything. Sure, maybe. First of all, I would disagree with that, but even if that were true, it’s not sensible. And that is exactly the problem. I don’t have to approach something with machine learning if I can get a better result with linear optimization. That’s one of those things where I would just say no. You also have to look at the intended use.

A1  
 Yes, exactly. And then with neural networks, with such large AI models, you lose topics like interpretability of the results, robustness of the results, because minimal changes in the data can suddenly trigger completely different internal processes. That is also something we could come back to later. Trustworthiness, we talked earlier about fairness and bias, which we don’t really cover on the label at the moment. Could you maybe say a bit about that? How do you see these areas?

I16  
 So trustworthiness has quite a few aspects, I think. First of all, it’s certainly a question of whether I can trust what the AI outputs, or whether there is potentially a problem, especially when I think of large models, that we have hallucination effects where the AI just makes things up, so to speak, or maybe I have a classic ML algorithm that simply does not have particularly good accuracy because it was not trained well, and then I suddenly have to realize that the results are wrong, and I can’t live with that. That’s one aspect. Another aspect, I think, is what happens with the data afterwards, and do these data lead to a responsibility, and does any human bear that responsibility, or does it suddenly belong to no one? And I think that is a much bigger area that is connected to it. This is something we already see very clearly when discussing whether AI content should be labeled or not. Can we even solve this automatically or not? What happens if we don’t, and who is responsible if something bad happens, for example if political campaigns are deliberately manipulated, or false information is deliberately spread? Those are also things that, I would say, strongly relate to this aspect of trustworthiness. But this is the application of these systems, and it is not something that is inherent to the systems themselves. So you couldn’t measure it just based on the systems.

A1  
 Yes, exactly. So a lot of that is difficult to reduce to metrics, and also fairness, for example, of an AI model is very hard to quantify. But of course it is something we think about in the context of the label, whether it can somehow be included. For example, one could say that the dataset was screened by experts for fairness or bias, and that such topics were consciously addressed in the curation and training of the model. Exactly, so we are also just a bit in a situation where this is an open question for us as well. How do we manage that? It is also important, in, well, we had a few conversations where people said, well, it’s a bit like a certificate when you have concrete values on it. Just like the EU energy label, it is also a form of guarantee for how this product works. And people trust it and say, okay, it has an energy efficiency class of B, that’s quite good, and A would be better. Is trustworthiness even given with such a label? Or would people question it? I don’t know, what do you think?

I16  
 I think you just have to communicate what the label does not say. So if a label, and I think this is actually very, very difficult, makes no statement about how fair a model is or how trustworthy a model is, how safe a model is, and you exclude that from the outset, then it can help at this point, but you basically need a warning label directly under the label stating that these things are not covered or considered. And then, psychologically, of course, there is again the not-listed effect, which tends to get overlooked. And that might make it a bit more difficult. Generally, I would say this is also the case with the NUTRI-Score, for example, where people say, oh, NUTRI-Score is nice and green, I will just take it, and then they don’t even look at the ingredients anymore. And I think you have to be careful a bit here and direct people to say, this is nice for an initial selection, but then look at the details, because there might be a lot happening there that is important for you.

A1  
 Do you get the feeling that if someone, I mean, if you leave these aspects out, they stay focused on these technical measurements, do you think, if you were to use it, I mean the software framework to generate these labels, that you could make it public on GitHub, because I somehow developed this as part of my PhD. You can just upload a table of measurements and it generates a label for all the models you have tested. Do you think people, if you showed it like that in your everyday life, would say yes, it says so on the label, so it must be correct? Or would they ask, where does this come from? Would they trust this information?

I16  
 I think yes. But that’s also because I wouldn’t show this label, for example, to a decision-maker, I would show it to a development engineer. Yes, because a decision-maker can’t do anything with the things that are currently measured by the label. And that is a problem. They would basically need a business value label.

A1  
 Yes, and cost-benefit.

I16  
 And that is of course not apparent here. A development engineer would say, ah, okay, here I see energy consumption and robustness and so on, I can work with that, I’ll look at it, and then of course you can still ask, yes, where does this come from? And then you can say, here is an open calculation algorithm, anyone can view it, anyone can inspect it, and the classification is based on that. I think that is something relatively trustworthy.

A1  
 I mean, of course, I have to say, it’s as trustworthy as people can just look at my software. And since I don’t have the measurements myself, only the reports, there’s also only limited room to trick things. So yes, my software is open, so in that sense it is trustworthy. Whether people trust it or not, doesn’t necessarily depend on that. But yes, probably without the lower part, it would probably be something that is almost more interesting at a decision-making level. Then it just says overall score, okay, fine. But as soon as it has this technical information, it might be something that could confuse or unsettle people.

I16  
 Yes, so here, of course, there are a lot of things listed that I would say, ah, okay, dataset used for training, hardware it runs on, software version it runs with, a decision-maker can’t do anything with that. On the other hand, if you leave out all these details, I think it becomes relatively difficult to assign an overall meaning to the label, because the A that is listed here is misleading, because if accuracy is really important to me, then this is definitely the wrong model.

A1  
 Yes, the idea would be that the person who develops the models in advance would do that given the priorities in the use case. And yes, then given this weighting, the overall scoring would be determined. We are almost done with the hour. So I would actually like to end the interview here. Thanks for your attention. I would send you the slides afterwards in case you want to look through them again or somehow…